Cole McIntosh

AI Engineer & Founder, Mellow AI

Dallas, TX · cole@staymellow.ai · colemcintosh.io · GitHub · LinkedIn

AI engineer and founder of Mellow AI building production agent systems, document extraction pipelines, and developer tools in Python, TypeScript, and Rust. Combines hands-on systems work—from fine-tuned small models and zero-dependency parsers to multi-provider LLM infrastructure—with end-to-end delivery that measurably reduces operational load.

TECHNICAL STACK

Languages	Python, TypeScript, Rust, SQL
AI & Agents	Agent harnesses, LLM evals & calibration, LoRA fine-tuning (Qwen), RAG, MCP, structured outputs
Frameworks & SDKs	Pydantic AI, Vercel AI SDK, LangChain / LangGraph, LiteLLM, Zod, Next.js, Ratatui
Data & Cloud	Custom OCR & PDF parsing, document forensics, OpenTelemetry & Langfuse, SQLite, AWS, Azure, GCP

EXPERIENCE

Founder · Mellow AI — 2025 – Present

- Re-architected the core document extraction platform and UI for a global law firm under strict data boundaries, modernizing ingestion and schema-validated pipelines.
- Engineered a custom OCR and extraction pipeline cutting processing cost by 95% and latency by 75%.
- Built and shipped Openextract (v1.0 on PyPI, plus a full TypeScript port)—a multimodal extraction toolkit supporting 10+ model providers with Pydantic/Zod validation.
- Fine-tuned and released mev (LoRA on Qwen3.5-4B), an open-weight structured decision model reaching 86.7% dev accuracy with calibrated confidence (ECE 0.038).

AI & Data Engineer · Pay Ready — 2023 – 2025

- Architected a general-purpose support agent harness steered by composable prompts and skills, enabling rapid rollout of new operational workflows without bespoke pipelines.
- Replaced a failing spam filter with a structured binary LLM classifier in live ticketing, cutting support agent workload by ~60%.
- Shipped a production GPT-4 Vision lease and contract extraction API over 250K+ documents (<1% human correction).
- Built an internal Slack RAG assistant adopted daily across teams in place of a bought tool, cutting manual effort by 60% and retrieval time by 75%.

Scaled Tooling Analyst · Otter / CloudKitchens (now Atoms) — 2022 – 2023

- Embedded with operations and engineering to build internal Python data pipelines scoring ~500GB/week.
- Designed and shipped an internal Jira ticketing chatbot from discovery to production, cutting ticket resolution time by 30%.

Business Analyst · RealPage — 2021 – 2022

- Built SQL data models and analytics dashboards supporting 75+ internal users, reducing manual reporting effort by 25%.

OPEN SOURCE

LiteLLM (LLM Gateway · Python) — SSO admin API/UI, GCP Model Armor guardrails, Langfuse OTel tracing, usage/health dashboards, new providers, caching/search cost tracking, and Ollama stream fixes.

Pydantic AI (Agent Framework · Python) — xAI search tools & attachment lifecycles, subagent X search fallback, OpenAI input token counting, UI URL sanitization, and provider host/timeout config.

smolagents (Hugging Face · Python) — Fixed Enum/metaclass sandbox execution, added AWS Bedrock API-key support, and refactored MultiStepAgent test suite.

SELECTED PROJECTS

TeraCode (Rust · Multi-agent TUI) — Local-first Rust TUI & bounded DAG scheduler over 6 coding-agent CLIs with git worktree isolation, binary-safe patching, SQLite history, and autonomy policies.

pdfmd (Rust · Zero-dep Engine) — Zero-dependency Rust PDF-to-Markdown engine from scratch (PDF parser, DEFLATE decoder, font & layout reconstruction) with CLI and Python bindings.

Openextract (Python · TypeScript) — Model-agnostic Python/TypeScript extraction toolkit turning documents and media into validated Pydantic/Zod objects across 10+ providers.

Liedar (TypeScript · Forensics) — Document tampering screening library, CLI, and browser UI combining metadata, copied-region detection, and model assessments for human review.

mev (LoRA · Qwen3.5-4B) — Open-weight Qwen3.5-4B LoRA decision model for 2–24 option structured state selection with calibrated confidence (Hugging Face).

CERTIFICATIONS

- RAG, LangChain, LangGraph & Semantic Search · DeepLearning.AI
- AWS Cloud Technical Essentials · Amazon
- Azure Applied AI · Microsoft
- Google Cloud Fundamentals · Google